A Methodological Approach for Web Sites Reengineering

Estiévenart, Fabrice; Francois, Aurore; Henrard, Jean; Hainaut, Jean-Luc

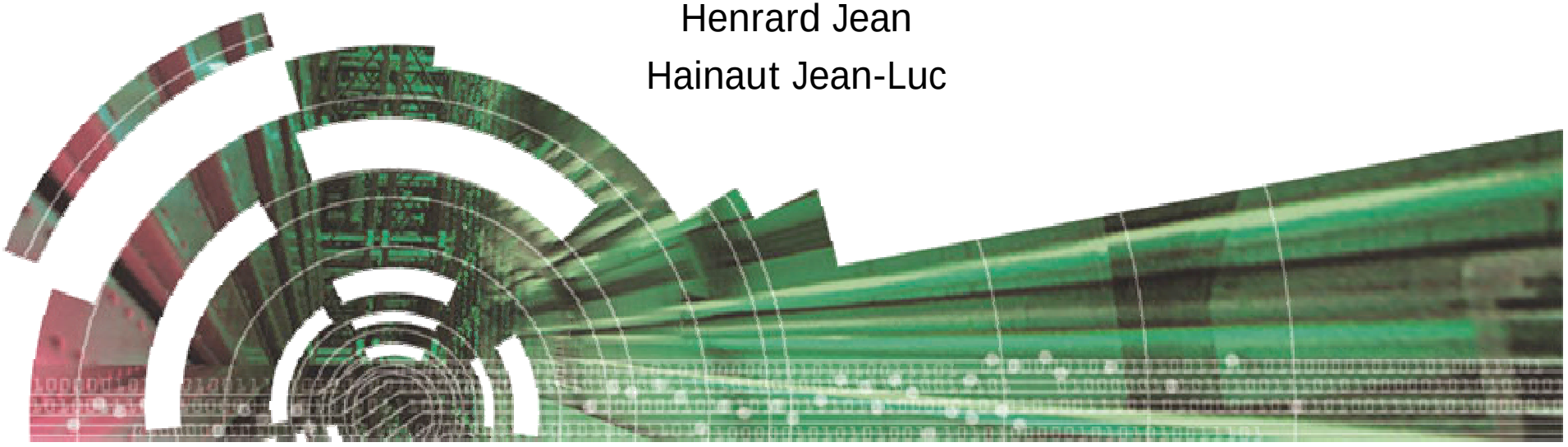
Publication date:
2003

cetic

Your connection to ICT research

A Methodological Approach for Web Sites Reengineering

Estiévenart Fabrice
François Aurore
Henrard Jean
Hainaut Jean-Luc



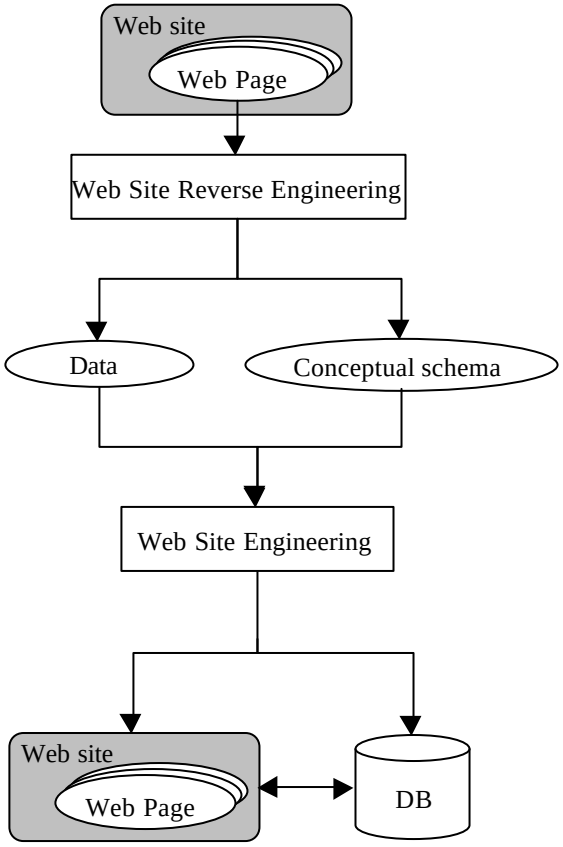
cetic
Your connection to ICT research

Context

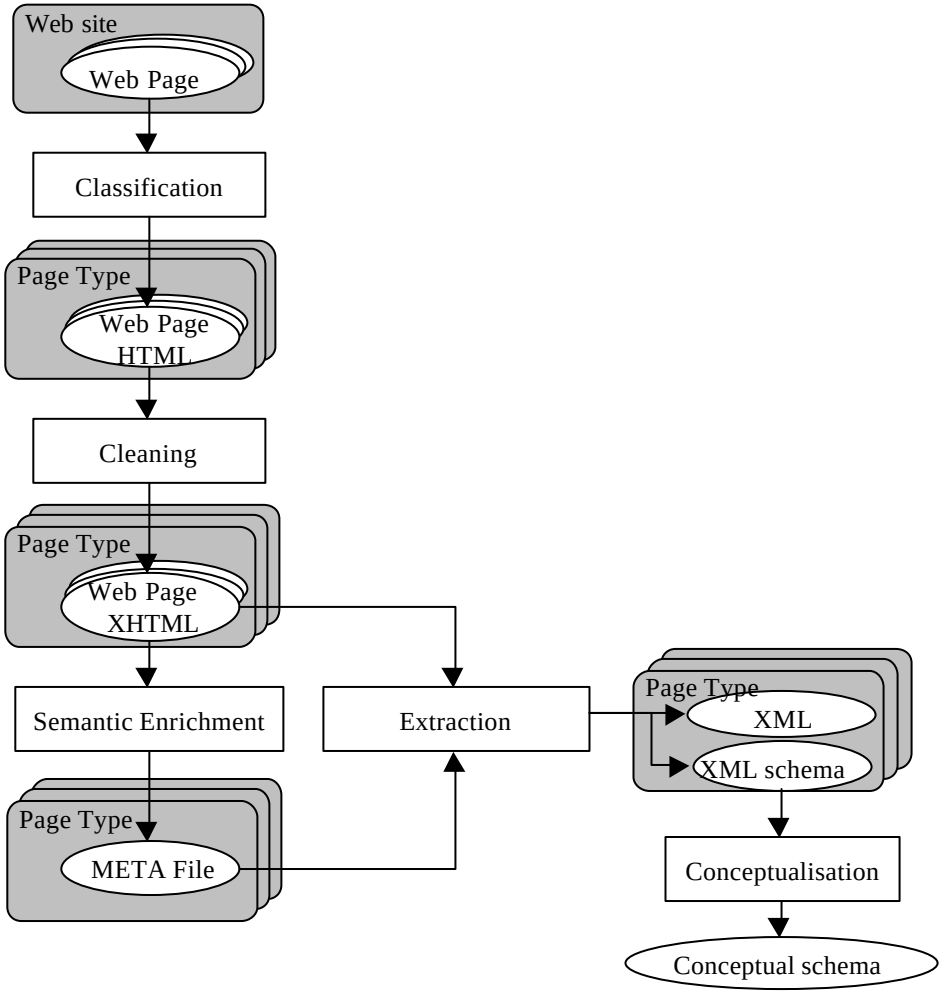
- Still many static web sites...
  - Advantage
    - easy to create
    - → for small web sites
  - Drawbacks
    - data and layout are mixed up
    - maintenance problems
    - → out-of-date or redundant information
    - → non-homogeneous design
  - Solution
    - DBMS + scripts (php, Perl,...)

Goals

- To provide methods and tools for web sites reengineering :

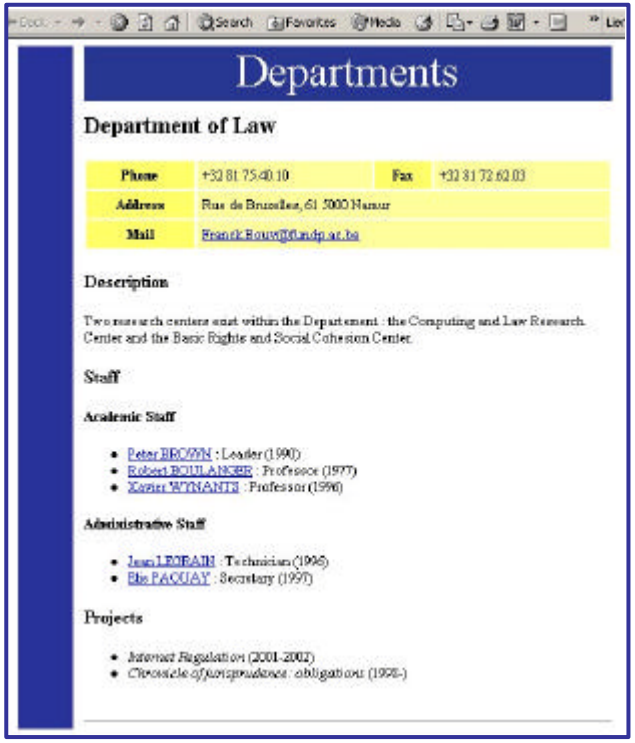


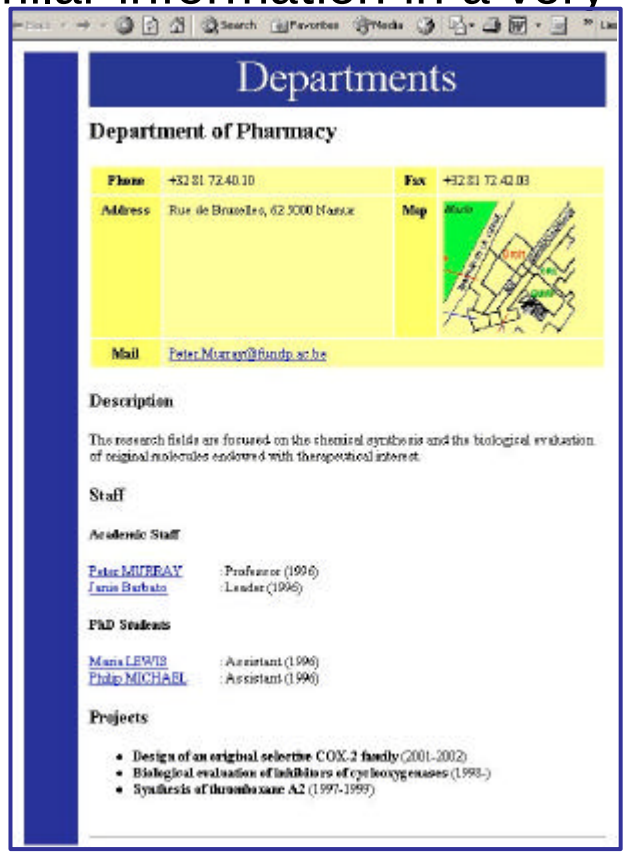
Method : overview



Method : step 1

- Pages classification
  - « Page type » = a set of pages relative to the same concept, that display very similar information in a very similar layout





Method : steps 2 and 3.1

- HTML cleaning
- Semantic enrichment
  - For each page type
    - Concepts identification and description on a sample page
      - « Concept » = a part of the HTML tree describing the layout, the structure and possibly the value of a certain reality
      - Ex : the concept « Phone Number »

```
<tr>
  <td align="middle" bgcolor="#FFFF66">
    <b>Phone :</b>
  </td >
  <td bgcolor="#FFFF99">+32 71 72.23.49</td>
</tr>
```

      - Ex : the concept « Address » composed of « Street » and « City »

```
<table width="100%">
  <tr><td><b>Address :<b></td></tr>
  <tr><td>Quality Street, 25</td></tr>
  <tr><td>London</td></tr>
</table>
```

Method : the META file

```
<HTMLDescription xmlns:meta="http://www.cetic.be/FR/CRAQ-DB.htm">
    <meta:element name="Department">
        <html>
            <head>...</head>
            <body>
             <table>
                <meta:element ref="DeptName"/>
                <meta:element ref="PhoneNumber"/>
                <meta:element ref="Address"/>
            </table>
            </body>
        </html>
    </meta:element>
    <meta:element name="PhoneNumber">
        <tr>
            <td align="middle" bgcolor="#FFFF66"><b>Phone :</b></td>
            <td bgcolor="#FFFF99"><meta:value/></td>
        </tr>
    </meta:element>
    ...
</HTMLDescription>
```

Method : step 3.2

- Application to other pages of the same type
  - there may be layout and structure differences between pages of the same type →a same concept may have several descriptions
  - **Example** : a layout difference

```
<tr>
    <td><b>Name :</b></td>
</tr>
```

```
<tr>
    <td><i>Name :</i></td>
</tr>
```

  - **Example** : a structure difference

```
<table width="100%">
    <tr><td>Address :</td></tr>
    <tr><td>Quality Street, 25</td></tr>
    <tr><td>London</td></tr>
</table>
```

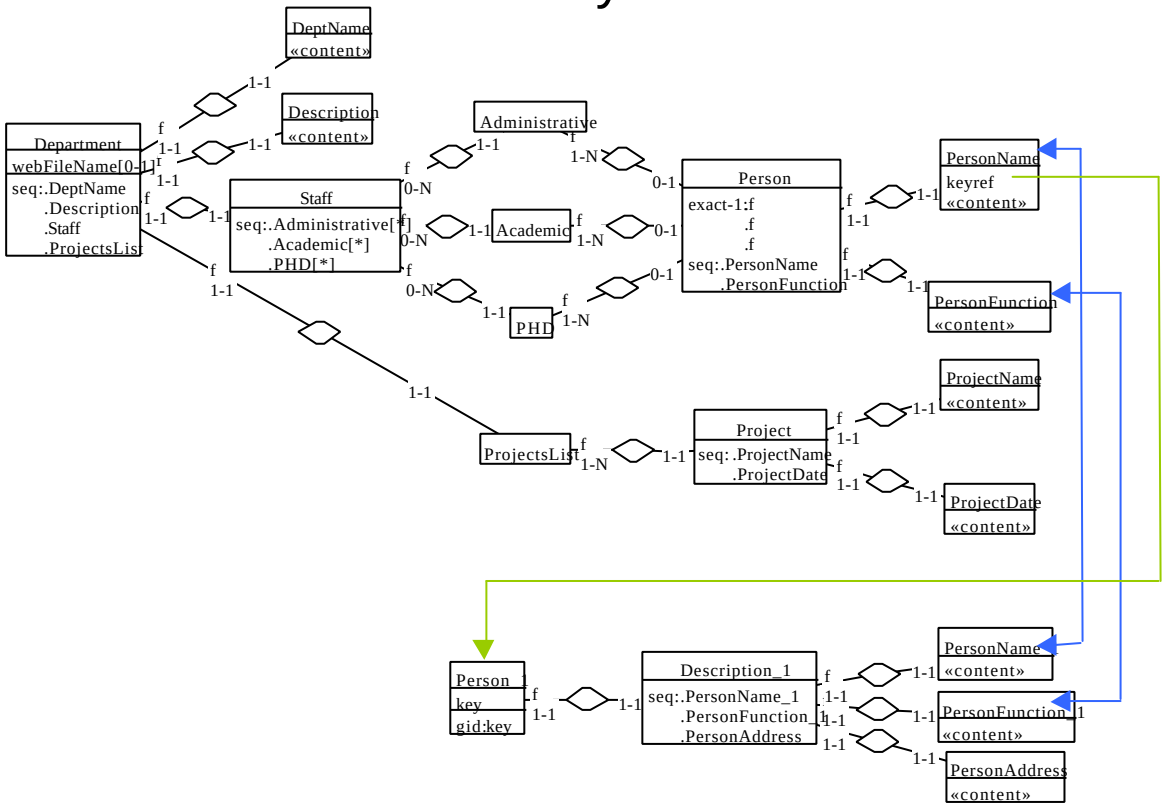
```
<table width="100%">
    <tr><td>Address :</td></tr>
    <tr><td>New York</td></tr>
    <tr><td>Main Street, 110</td></tr>
</table>
```

Method : step 4

- Data and schema extraction
  - Data extraction
    - Web pages + META file → XML document
  - Data structure extraction
    - META file → XML schema

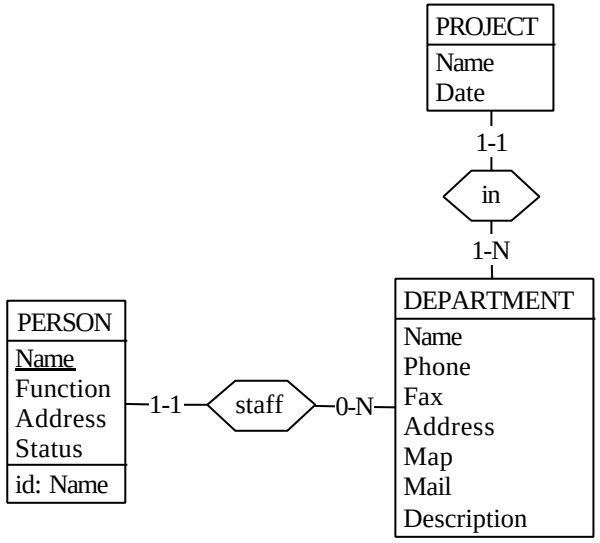
Method : step 5.1

- Schema integration
  - To discover relationships between concepts
  - To detect redundancy



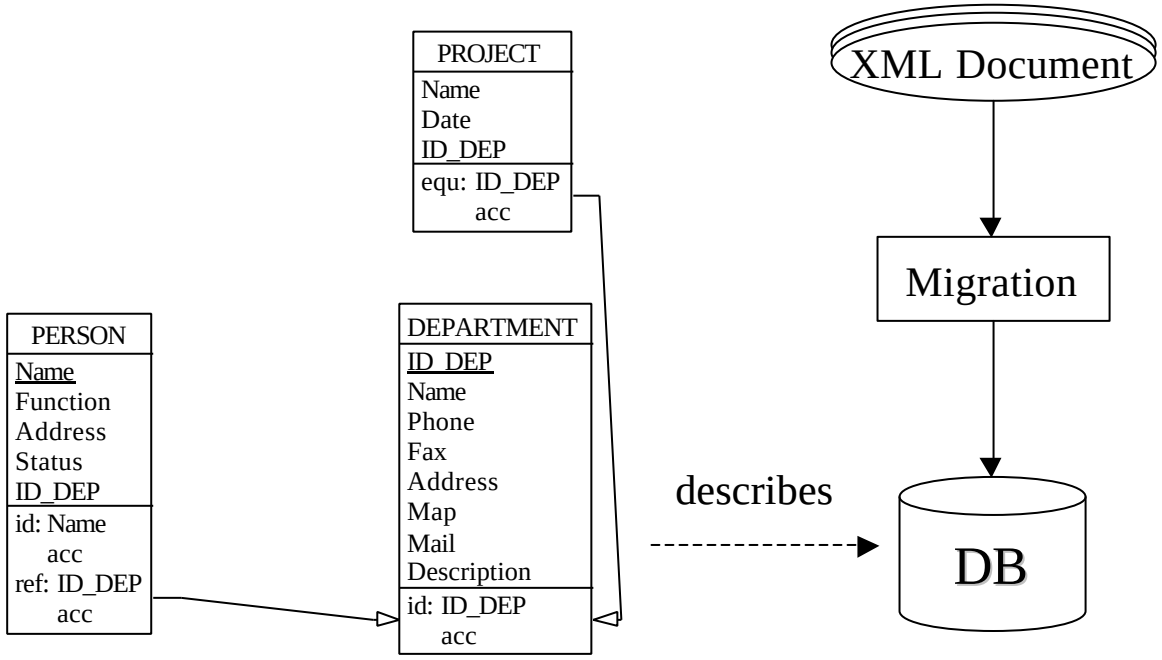
Method : step 5.2

- Schema conceptualisation



Web Site Engineering

- Database engineering + Data migration



cetic
Your connection to ICT research

Tools

- HTML cleaning
  - Tidy
- Semantic enrichment
  - XML editor or the semantic browser (based on Mozilla) to edit/generate the META file
- Data and schema extraction
  - XML parsers (Java DOM)
  - *pageType.extractSchema(METAfile)* → *XMLSchema*
  - *pageType.extractData(HTML*, METAfile)* → *XML*
- Schemas integration/conceptualisation and database engineering
  - CASE tool DB-Main

cetic
Your connection to ICT research

Conclusion and future work

- A method and tools to extract from a web site data and their structure
- Difficulty : enormous diversity of web pages structures and layouts to represent the same reality
- Future work
  - test on real-size web sites
  - refine the semantic enrichment step
    - improve GUI
    - automation/assistance based on heuristics